

Beyond Semantic Image Segmentation : Exploring Efficient Inference in Video

Subarna Tripathi¹, Serge Belongie², Truong Nguyen¹

¹University of California San Diego. ²Cornell NYC Tech.

Deep convolutional neural networks (DCNNs) trained on a large number of images with pixel-level annotations or a combination of strongly labeled and weakly-labeled images have recently been the state-of-the-art in semantic image segmentation with significant performance improvement.

However, due to the very invariance properties that make DCNNs good for high level tasks such as classification, visual delineation capacities for deep learning techniques are limited. Recent approaches address this problem with Conditional Random Field (CRF) based graphical model in two ways:

1. Adding a post-processing step of CRF-based probabilistic graphical model for the pixel-level classification [3, 8].
2. Integrating the graphical model as a part of the CNN to make the end-to-end learning with the usual back-propagation possible without the need of post-processing [11].

In either case, the final pixel-level classification accuracy and efficiency remain highly dependent on the inference step of the image-based CRF [5] involved where fast approximate MPM inference is performed using cross bilateral filtering techniques within a mean-field approximation framework.

Alvarez *et al.* [1] demonstrates that performing inference on all test images at once in a dense CRF yields better results than inferring one image at a time without additional computation cost compared to performing segmentation sequentially on individual images. It is to be noted that the dense CRF [5] achieves good results with only unary and pairwise terms. This fully-connected pair-wise model is more expressive than its 4 or 8-connected random field counter-parts. Yet, it lacks the ability to handle high-order terms. Models [4, 6, 7] using higher-order terms such as label consistency over large regions (pattern-based potentials) and relations of global co-occurrence potentials, are shown to be more expressive for object class segmentation task. Filter-based inference for those higher-order terms is formulated in [10] which enables significant speed-up compared to those graph-based methods [4, 6, 7]. Yet, it needs to consider temporal consistency when applied in co-segmentation or video semantic segmentation.

We explore the efficiency of the CRF inference module beyond image level semantic segmentation. The key idea is to combine the best of two worlds of semantic co-labeling and exploiting more expressive models. Similar to [1] our formulation enables us perform inference over ten thousand images within seconds. On the other hand, it can handle higher-order clique potentials similar to [10] in terms of region-level label consistency and context in terms of co-occurrences. We follow the mean-field updates for higher order potentials similar to [10] and extend the spatial smoothness and appearance kernels [5] to address video data inspired by [1]; thus making the system amenable to perform video semantic segmentation most effectively.

Figure 1 shows some qualitative results of semantic segmentation in Camvid video dataset [2]. In this particular experiment, we used the TextonBoost [9] unary potentials for easy comparison with other recent methods. Video-Level Dense-CRF [1] shows improved temporal consistency over frame-level operation [5] (previous row) without additional time overhead. For, pattern-based potentials, we use three different superpixel segmentations by varying parameters of the meanshift algorithm. Frame-level Dense-CRF with this P^n -Potts model [10] almost achieves similar quality as of previous graph-cut based slow inference method [6], but lacks temporal consistency. The proposed video-level Dense-CRF with P^n -Potts model shows improved temporal consistency over the frame-level operation (previous row) without additional time-overhead. Video-Level dense CRF [1] and the proposed method perform inference on 50 frames at once. On this dataset, with TextonBoost unaries our proposed method achieves 8% more accuracy than [1] by virtue of P^n -Potts model and 1.5% more accuracy over [10] without additional time overhead by virtue of co-labeling. CNN feature classification yields improved unary potentials compared to the unaries

provided by TextonBoost. Analyzing the final video semantic segmentation accuracy using CNN based unaries and proposed dense-CRF with P^n -Potts model remains our future work.

- [1] J.M. Alvarez, M. Salzmann, and N. Barnes. Large-scale semantic co-labeling of image sets. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 501–508, March 2014. doi: 10.1109/WACV.2014.6836060.
- [2] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recogn. Lett.*, 30(2):88–97, January 2009. ISSN 0167-8655. doi: 10.1016/j.patrec.2008.04.005. URL <http://dx.doi.org/10.1016/j.patrec.2008.04.005>.
- [3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proc. ICLR*, 2015.
- [4] Pushmeet Kohli, Lubor Ladicky, and Philip H.S. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.
- [5] Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *International Conference on Machine Learning (ICML)*, pages 513–521, 2013.
- [6] L. Ladicky, C. Russell, P. Kohli, and P.H.S. Torr. Associative hierarchical crfs for object class image segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 739–746, Sept 2009.
- [7] Lubor Ladicky, Paul Sturgess, Karteek Alahari, Chris Russell, and Philip H.S. Torr. What, where and how many? combining object detectors and crfs. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision, 2010 IEEE 12th International Conference on*, volume 6314 of *Lecture Notes in Computer Science*.
- [8] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. *CoRR*, abs/1502.02734, 2015. URL <http://arxiv.org/abs/1502.02734>.
- [9] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.
- [10] Vibhav Vineet, Jonathan Warrell, and Philip HS Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *International Journal of Computer Vision*, 110(3):290–307, 2014.
- [11] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. *CoRR*, abs/1502.03240, 2015. URL <http://arxiv.org/abs/1502.03240>.

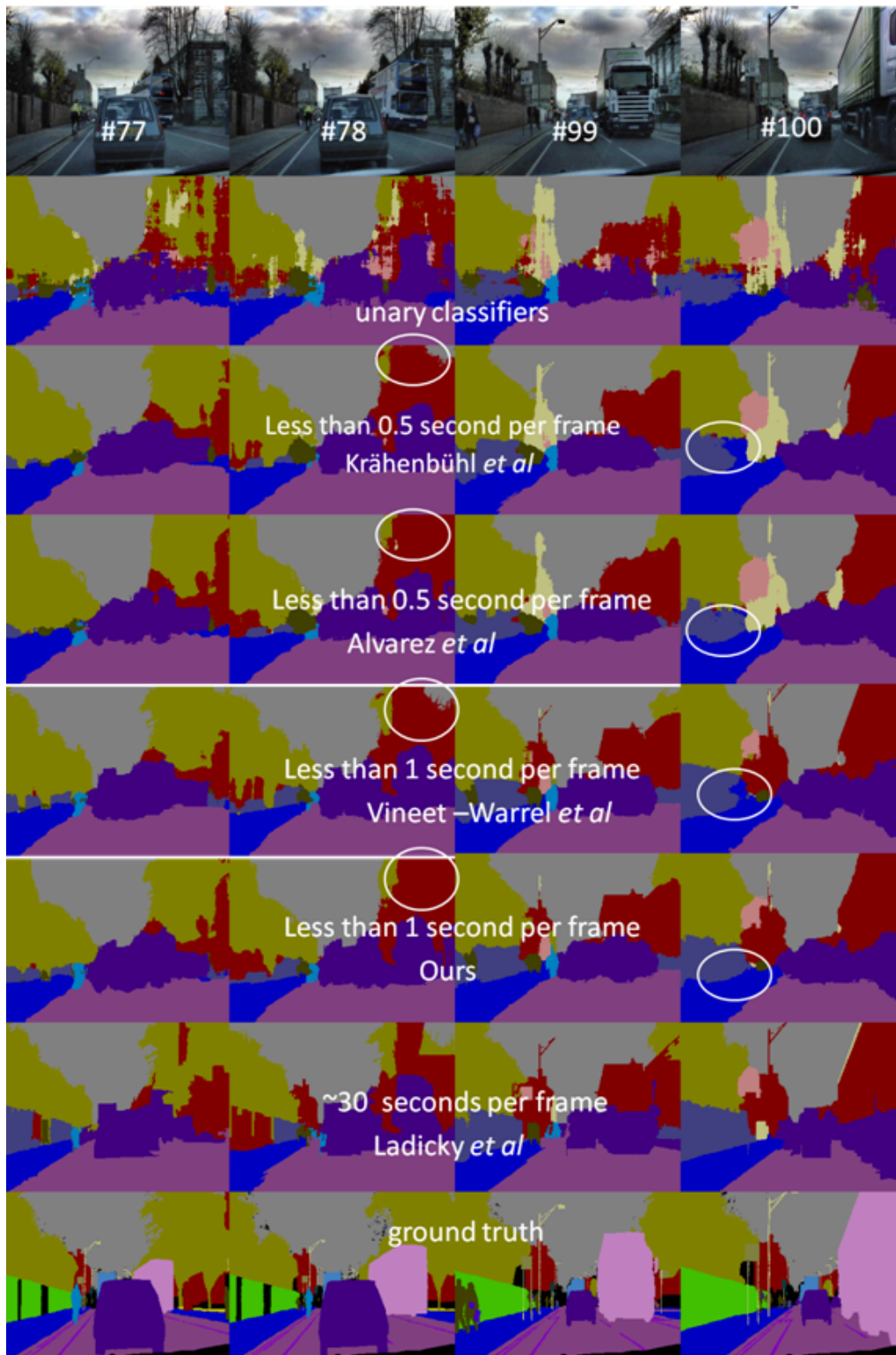


Figure 1: Qualitative results on Camvid dataset [2]. From top to Bottom : Input frames, Unary potentials from TextonBoost classifier scores [9]; Frame-level Dense-CRF [5]; Video-Level Dense-CRF [1] shows improved temporal consistency over frame-level operation (previous row) without additional time overhead; frame-level Dense-CRF with P^n -Potts model [10]; Proposed video-level Dense-CRF with P^n -Potts Model shows improved temporal consistency over the frame-level operation (previous row) without additional time-overhead; frame-level Graph-cut based slow inference with P^n -Potts Model and the Ground truth levels.